*DOI:*
10.1002/joc.6215

*Document Version*
Peer reviewed version

# Deriving optimal weather pattern definitions for the representation of precipitation variability over India

*Weather patterns and precipitation variability over India*

Robert Neal[1,*], Joanne Robbins[1], Rutger Dankers[1], Ashis Mitra[2], A. Jayakumar[2], E. N. Rajagopal[2], George Adamson[3]

1. Met Office, FitzRoy Road, Exeter, Devon, EX1 3PB, UK (***Contact:** *robert.neal@metoffice.gov.uk*; +44(0) 330 135 2166)
2. National Centre for Medium-Range Weather Forecasts (NCMRWF), Noida, India
3. Kings College London, UK

**Short abstract**

Cluster analysis is used to generate a set of 30 weather patterns which represent variability within different phases of the Indian climate. Weather pattern variants are evident within the active monsoon, break monsoon, retreating monsoon and western disturbances. These weather pattern variants are useful when it comes to identifying periods most susceptible to high impact weather within a large-scale regime, such as identifying the most flood prone periods within the active monsoon, and have potentially many forecasting applications.

**Abstract**

This study utilises cluster analysis to produce sets of weather patterns for the Indian subcontinent. These patterns have been developed with future applications in mind; specifically relating to the occurrence of high impact weather and meteorologically-induced hazards such as landslides. The

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/joc.6215

weather patterns are also suited for use within probabilistic medium- to long-range weather pattern forecasting tools driven by ensemble prediction systems. A total of 192 sets of weather patterns have been generated by varying the parameter which is clustered, the spatial domain and the number of weather patterns. Non-hierarchical k-means clustering was applied to daily 1200 UTC ERA-Interim reanalysis data between 1979 and 2016 using pressure at mean sea level (PMSL) and *u*- and *v*-component winds at 10-m, 925-hPa and 850-hPa. The resultant weather pattern sets (clusters) were analysed for their ability to represent the main climatic precipitation patterns over India using the explained variation score. Weather patterns generated using 850-hPa winds are among the most representative, with 30 patterns being enough to represent variability within different phases of the Indian climate. For example, several weather pattern variants are evident within the active monsoon, break monsoon and retreating monsoon. There are also several variants of weather patterns susceptible to western disturbances. These weather pattern variants are useful when it comes to identifying periods most susceptible to high impact weather within a large-scale regime, such as identifying the most flood prone periods within the active monsoon. They hence have potentially many forecasting applications.

**Keywords:** Weather patterns, cluster analysis, India, monsoon, forecasts

## 1. Introduction

This study uses cluster analysis to define a set of representative weather patterns for the Indian subcontinent. These weather patterns represent all the main monsoonal and non-monsoonal circulation types which occur throughout the year and are tested against their ability to explain precipitation variability. A weather pattern can be described as one of many circulation types over a

defined region (e.g. the Indian subcontinent or smaller regional area) which differs in its characteristics from other weather patterns over the same region and can vary on a daily basis. The term weather regimes can also be used to describe a defined circulation type, where weather regimes are typically larger in scale, fewer in number, and persist for more days than weather patterns. Large-scale weather regimes over India follow a relatively predictable evolution throughout the course of the year, driven predominantly by the onset and retreat of the Asian summer monsoon (Annamalai *et al.*, 1999; Islam *et al.*, 2018). This study will investigate circulation variability within each of these large-scale weather regimes, with the aim of identifying the sub-weekly weather patterns responsible for the within-regime precipitation variability.

Cluster analysis is an established and effective statistical method used across a range of scientific disciplines for identifying natural groups or clusters within a sample of data (Andenberg, 1973). It is particularly suited to clustering weather and climate data with a range of applications; for example, clustering air trajectories over the UK and relating the resulting groups to air quality (Dorling *et al.*, 1992); clustering 500-hPa pressure fields to reproduce four dominant large-scale winter weather regimes over Europe (Bao and Wallace, 2015) and clustering reanalysis data for a range of parameters over the United States and relating the resultant patterns to historical changes in precipitation (Prein *et al.*, 2016). Cluster analysis has also been used within weather forecasting applications. For example, Molteni *et al.* (1996) show how clustering of Numerical Weather Prediction (NWP) ensemble members was used in early ensemble forecast products at the European Centre for Medium Range Weather Forecasts (ECMWF). More recently, clustering of NWP ensemble members has been achieved by objectively assigning forecast members to the closest matching weather pattern definition (generated through cluster analysis), enabling probabilistic weather pattern forecasts to be produced (Neal *et al.*, 2016). Clustering NWP ensemble forecast members in this way has the benefit that

assumptions can be made about the weather conditions or weather impacts expected based on the characteristics associated with each forecast weather pattern. This enables probabilistic weather pattern forecasts to be used within specific end-user applications, for example relating forecasts of particular weather patterns to the risk of coastal flooding (Neal *et al.* 2018).

This study applies a non-hierarchical k-means clustering algorithm to daily gridded reanalysis fields from ERA-Interim (Dee *et al.*, 2011) covering the period from 1979 to 2016. In total, 192 sets of weather patterns were produced by varying the parameter clustered, the domain used for clustering, and the number of clusters produced. Emphasis was placed on the ability of different weather pattern sets to represent variability in precipitation, which is one of the most important meteorological parameters to consider within the Indian climate for a number of reasons. India is heavily reliant on rainfall for agriculture (Gadgil and Gadgil, 2006). It also experiences significant impacts from flooding, such as the Kerala floods in August 2018 (Mishra and Shah, 2018), and rainfall induced hazards, such as the numerous landslides which occurred in the state of Uttarakhand (northern India, on the western border with Nepal) during June 2013 (Martha *et al.*, 2015). Many of the weather pattern domains tested in this study focus particularly on two landslide prone regions of India: Darjeeling District (a Himalayan region in the northern most part of the state of West Bengal, north-east India) and Nilgiris District (a hilly region in the north-western most part of the state of Tamil Nadu, southern India).

Previous studies to use cluster analysis for the generation of European weather patterns focus on clustering gridded pressure fields. These include Fereday *et al.* (2008), who created a set of 10 weather regimes for each two-month period throughout the year; Ferranti *et al.* (2015), who created a set of four weather regimes which are valid over the whole year; and Neal *et al.* (2016), who created a

set of 30 weather patterns centred over the UK, also valid over the whole year. Several previous studies over India have focussed on identifying dominant rainfall patterns by clustering rainfall observations during the monsoon season (June though to September). For example, Bedi and Bindra (1980) identified 15 dominant rainfall patterns which represent 72% of the total variance in monsoon rainfall by applying principal components analysis (PCA) to rainfall data from 70 gauges covering India over a 60 year period. Gadgil *et al.* (1993) identified seven dominant rainfall patterns which represent between 50 and 60% of the total variance in monsoon rainfall by applying PCA to rainfall data from 138 gauges covering India over an 87 year period; they also identified between 11 and 31 coherent rainfall zones throughout the monsoon period which are useful for understanding rainfall variability at different spatial scales. Lastly, Krishnamurthy and Shukla (2000) applied PCA to rainfall data from 3700 gauges covering India over a 70 year period to analyse intraseasonal and interannual variations in monsoon rainfall. In results relevant to this study they found that the main intraseasonal variability of monsoon rainfall is characterised by the occurrence of active and break periods.

More recently, studies over India have looked at the large-scale circulation and then related this back to rainfall. For example, Chattopadhyay *et al.* (2008) investigated the summer monsoon intraseasonal oscillation (ISO) using a pattern recognition technique known as self-organising maps (SOMs). This technique used six daily large-scale circulation indices derived from the NCEP-NCAR reanalysis (Kalnay *et al.*, 1996). A set of 81 and 9 weather patterns (nodes) were identified using a 9 × 9 and 3 × 3 lattice respectively. The resultant 9 patterns included three active monsoon types, three break monsoon types and three normal state types. It was shown that the weather patterns related very strongly to the rainfall ISO even though the SOM input indices were representative of the large-scale circulation. Islam *et al.* (2018) also applied the SOMs technique to NCEP-NCAR reanalysis data, but this time for identifying the full seasonal cycle across South Asia. SOMs were applied separately to

five near-surface variables (wind, precipitation, sea-level pressure, relative humidity and temperature) using a preferred 9 × 9 lattice. The resulting 81 weather patterns (nodes) were merged into 8 groups, which represent the four dominant seasons (winter, pre-monsoon, monsoon and post-monsoon) as well as the four transitions between these seasons.

This paper will use clustering methods which have already been applied and proven successful over the temperate climate of Europe to the largely tropical climate of India. In contrast to previous studies, it presents a new set of weather patterns which are responsible for precipitation variability throughout the whole year and not just during the summer monsoon, thus capturing all precipitation variability over India. It will also demonstrate how these new weather patterns could help open up opportunities for future research and forecasting applications, such as helping to understand the synoptic-scale driving forces behind the occurrence of high impact weather (e.g. intense thunderstorms) or the occurrence of specific meteorologically driven hazards (e.g. landslides) and for use within probabilistic medium- to long-range weather pattern forecasting tools.

## 2. Data and methodology

### *2.1. Clustering method*

The weather pattern clusters were produced by applying a non-hierarchical k-means clustering algorithm provided by SciPy (Jones *et al.*, 2001), which is an open source scientific tools package using Python code. The algorithm attempts to minimize the Euclidian distance between observations (daily gridded reanalysis fields) and centroids (mean fields for each weather pattern). As discussed in the introduction, this clustering method has been successfully applied in many previous studies for

identifying circulation types (particularly over Europe and the United States). Other methods such as PCA and SOMs have also been shown to be successful methods within similar applications and could equally have been used. However, to a large extent, the specific clustering method chosen is less relevant assuming the resultant weather patterns represent enough variability in the main parameter of interest (e.g. precipitation).

In total, 192 sets of weather patterns were produced by changing three variables (4 parameters × 12 spatial domains × 4 cluster numbers), which are explained in the following paragraphs. The final set of patterns was then arrived at by an iterative process, in which each set was evaluated for its ability to represent the full range of circulation types over India along with key precipitation patterns (Section 2.2).

Firstly, the data clustered needs to be specified. The weather pattern sets were generated by clustering ERA-Interim reanalysis data (Dee *et al.*, 2011) which was processed on a 1 × 1 degree resolution grid. Daily fields at 1200 UTC were clustered between 1979 and 2016. Several parameters were clustered before examining which parameter produces the most representative set of weather patterns. The parameters clustered were pressure at mean sea-level (PMSL), 10-m wind vectors, 925-hPa wind vectors and 850-hPa wind vectors. PMSL was considered important due to the formation of low pressure over land during the active monsoon when the pressure gradient between the Arabian Sea and land is well organised. Wind fields were also considered important due to India's climate being heavily influenced by changes in the direction of the prevailing wind between monsoon and non-monsoon periods. A humidity variable representing atmospheric moisture capacity was not considered at this stage, although is also important for India's climate. However, moisture information

is implicitly included within the weather patterns generated from the wind vectors due to moisture being directly related to wind direction in this part of the world.

Secondly, the spatial domain used in the clustering needs to be specified. In this study 12 domains were tested (Figure 1). This includes three domains covering the entire Indian subcontinent; five domains centred on Darjeeling District and four domains centred on Nilgiris District. The smaller domain sets centred on Darjeeling and Nilgiris Districts were used to test the assumption that weather patterns generated using smaller domains have a better relationship with precipitation variability, as already demonstrated by studies over Europe (e.g., Beck *et al.* (2013) and Neal *et al.* (2016)). These particular districts were chosen as they are both highly sensitive to hydro-meteorological hazards including floods and landslides, thus potential application sites for the weather patterns generated. Darjeeling and Nilgiris Districts also represent different precipitation climates, with Darjeeling receiving a large proportion of its rainfall from break and pre/post monsoon patterns, whereas Nilgiris receives a large proportion of its rainfall from active and retreating monsoon patterns.

Finally, the number of resultant clusters (weather patterns) needs to be decided. The choice of number of patterns is, to a certain extent, a subjective compromise. With too few patterns, the variability of precipitation within each pattern is too large. Too few patterns also means that any weather impact of interest (such as landslides) are likely to fall under the same pattern which will typically persist for many days before transitioning. This makes it difficult to identify high-risk periods within a persisting weather pattern. In contrast, too many patterns means that neighbouring pattern centroids begin to look similar. As a result, any weather impact of interest is likely to be distributed among several similar types. Also, weather impacts are typically associated with severe weather, which is rare in its occurrence. Therefore, the small sample of weather impacts will be distributed among multiple

similar-looking weather patterns, thus making it difficult to identify high-risk periods within a forecasting application with any confidence. In this study, clustering was carried out using 10, 20, 30 and 40 weather patterns. The resultant weather patterns are ordered according to their observed frequencies during the clustering period, with the lower number patterns occurring most often and the higher numbered patterns occurring least often.

The clustering method has two outputs. Firstly, daily historical weather pattern classifications (i.e. weather pattern numbers) for each day within the gridded input data set are provided. These classifications can be used to relate weather patterns with a parameter of interest where historical records exist, such as weather observations or cases of weather induced hazards (e.g. landslides). Assuming a relationship exists, it then becomes possible to make assumptions as to the likely consequences given the occurrence of a particular weather pattern. Secondly, the centroids (definitions) for each weather pattern are output on a grid. These are calculated using the daily historical weather pattern classifications by taking a mean of the input ERA-Interim reanalysis fields for each weather pattern. These two outputs for the optimal set of weather patterns introduced in Sections 3 and 4 can be downloaded from the PANGAEA online data repository (Neal *et al.*, 2019).

### *2.2. Evaluation of weather patterns*

All weather pattern sets were objectively analysed for their ability to represent precipitation variability at 10 locations across India (Table 1 and Figure 1), which were chosen as they represent a range of precipitation climates, are spatially distributed and occur at different heights above sea-level. The Explained Variation (EV) score given in Equation 1 (from Beck *et al.*, 2013) is used to measure how well a daily precipitation series at a given location can be reconstructed based on knowledge only of

the weather pattern classification. EV results are then analysed across all locations for each weather pattern set. We try to ensure that as many locations as possible have the best EV scores for the preferred set of weather patterns.

$$EV = \frac{\sum_{k=1}^{K} N_k(\overline{a}_k - \overline{a})^2}{\sum_{i=1}^{N}(a_i - \overline{a})^2} \quad (1)$$

In Equation 1, *N* is the number of days across the whole sample (irrespective of weather pattern), *K* is the number of weather patterns, $N_k$ is the number of days in a given weather pattern, $a_i$ is the observed rainfall on a given day, $\overline{a}$ is the mean rainfall across the whole sample (irrespective of weather pattern) and $\overline{a}_k$ is the mean rainfall for a given weather pattern. The score ranges between zero and one, with the higher the score the better. In the context of this study, a score of zero means that the variance in precipitation among weather patterns is the same as the variance in precipitation among the full sample. In contrast, a score of one means that the variance in precipitation among weather patterns is orthogonal to the variance in precipitation among the full sample. Other comparison measures could also have been used, such as the Pseudo-F statistic and Silhouette index, which are among six measures introduced in Beck and Philipp (2010) for identifying optimal circulation type classifications within Europe. However, EV provides among one of the best overall measures due to its simple comparison of within type variability to the full sample variability and has been consistently used across a number of studies (e.g., Beck and Philipp (2010), Casado and Pastor (2013), Beck *et al.* (2013) and Neal *et al.* (2016)).The location specific precipitation time series used within the EV calculations (done separately for the 10 locations in Table 1) came from the Indian Meteorological Department (IMD) who produced a 0.25 × 0.25 degree resolution gridded precipitation observation data set covering the whole of India (Pai *et al.*, 2014). This data set is formed from an interpolation of 6955 land-based rain gauges and should provide an improved

representation of orographically enhanced precipitation compared to ERA-Interim (Dee *et al.*, 2011). The high resolution daily rainfall fields from IMD were neighbourhood post-processed before being used in this study. This method was applied so that rainfall totals at each grid-point represent a mean of all values within a two grid-cell square neighbourhood (5 × 5 grid). This has the effect of smoothing out some of the more localised spatial variability in rainfall totals found over areas with varying orography (such as in the Darjeeling and Nilgiris Districts), whilst retaining information on extremes. The smoother rainfall field removes any localised spatial sensitivity when selecting data for a specific location (for the 10 locations in Table 1), which are then used in the EV calculations.

## 3. Results Part 1 – Identifying the optimal set of weather patterns

Choosing an optimal set of weather patterns involves identifying the best combination of parameter (4 parameters tested), domain (12 domains tested) and number of resultant patterns (4 pattern sets tested). This involved running the clustering method 192 times for all possible combinations. The EV score was then calculated for all 192 weather pattern sets to help identify the optimal set of weather patterns for precipitation variability across India. The three clustering variables (parameter, domain and number of patterns) will now be addressed separately.

### *3.1. Clustering parameter*

The four clustering parameters were tested were PMSL, 10-m *u*- and *v*-component winds, 925-hPa *u*- and *v*-component winds and 850-hPa *u*- and *v*-component winds. Focus was placed on wind due to India's climate being heavily influenced by changes in the direction of the prevailing wind. For example, wind direction reverses from a dry north-westerly in winter to a moist south-westerly

in summer. Such changes in wind direction are triggered by the formation and decay of heat lows over land. For example, low pressure forms over the Indian subcontinent as the land mass heats up in early summer. As a result, moisture laden winds flow from the relative cold Arabian Sea (where there is high pressure) towards the hot and dry Indian land mass (where there is low pressure). This also suggests that PMSL may be a sensible parameter to investigate, but possibly only relevant during the active monsoon when the pressure gradient is well organised. Results show that patterns generated using PMSL produce the worst EV scores for precipitation for 9 of the 10 geographically varying locations across India shown in Figure 1. Only in Darjeeling does PMSL rank 2$^{nd}$ after 850-hPa wind (Table 1) when considering precipitation variability among a set of 30 weather patterns generated using the India Domain 2 (Figure 1). This suggests that precipitation variability in Darjeeling has an unusually close relationship with variability in PMSL. Weather patterns generated using *u*- and *v*-component winds at 850-hPa have the best overall relationship with precipitation variability, closely followed by 925-hPa winds, with 10-m winds coming in 3$^{rd}$ place overall. The results in Table 1 are consistent with what was seen across most other pattern sets and domains.

### *3.2. Clustering domain*

Three domains were tested which cover the Indian subcontinent (Figure 1). These have varying sizes with India Domain 1 being the smallest (excluding the northern most parts of India) and India Domain 3 being the largest (extending as far north and east as Afghanistan and as far west as Myanmar). In addition, between 4 and 5 smaller domains were tested for both Nilgiris and Darjeeling (Figure 1), which are geographically located at opposite sides of India and also have different precipitation climatologies. EV scores for precipitation variability in Central Nilgiris and Darjeeling are presented in Figures 2 and 3 respectively, where results are restricted to weather patterns generated using *u*- and

*v*-component winds at 850-hPa (which has already been identified as the best parameter to use for generating the weather patterns in Section 3.1). Figures 2 and 3 also include a breakdown of EV results for different numbers of weather patterns, with analysis of this final clustering variable covered in Section 3.3.

EV scores for Central Nilgiris (Figure 2) show that precipitation variability here is best represented by weather patterns generated using the Central Nilgiris domains, with the smaller domains generally performing best (Nilgiris Domain 1 is better than Nilgiris Domain 4). This is closely followed by the Indian domains where the largest Indian domain demonstrates a small advantage over the smaller two India domains, although results are slightly dependant on the number of weather patterns. The Darjeeling domains unsurprisingly all have very low EV scores, with the largest of the Darjeeling domains performing best – possibly because it captures more of Southern India where Nilgiris District is located.

EV scores for Darjeeling (Figure 3) show that precipitation variability here is best represented by weather patterns generated using the three Indian domains – which all have very similar EV scores. Surprisingly, the small weather pattern domains centred on Nilgiris District have EV scores which are equally as good as the EV scores for the small weather pattern domains centred on Darjeeling itself, but still not as good as the Indian domains. This suggests that weather patterns defined over Southern India have a good relationship with precipitation variability in Northeast India. Overall, the Indian domains provide a good compromise for Central Nilgiris (providing the second best clustering domain) and are the best domain for Darjeeling. In addition, the Indian domains always have at least the second best EV scores (from the three domain groupings) when considering precipitation variability at all 10 locations shown in Figure 1. The final set of weather patterns are intended for use

within a range of forecasting applications which cover the full spatial extent of India. Therefore, one set of weather patterns relevant for the whole of India is preferable from a practical point of view as it would allow one weather pattern forecasting system to be used by multiple applications. There is very little difference in EV scores for the three Indian domains; therefore India Domain 2 has been chosen due to it being the smallest domain which covers the whole of India.

*3.3. Number of weather patterns*

The final variable within the clustering process is the number of weather patterns, which was set to 10, 20, 30 and 40. EV scores (Figures 2 and 3) reveal that the larger the number of patterns the better the ability of the weather patterns to explain variability in precipitation. This result is consistent with Beck and Philipp (2010) and is to be expected given that the more a precipitation time series is split up then the more likely it is that the sample means between each group will differ. The EV scores (Figures 2 and 3) suggest that the benefit of increasing the number of weather patterns starts to tail off at around 40 patterns, with the difference in EV scores between 30 and 40 patterns being very similar in many cases, relative to the difference in EV scores between 20 and 30 patterns. However, we did not test beyond 40 patterns, so further research would be required to test this assumption. In the end, 30 patterns was chosen with the objective analysis showing that 30 patterns is always in the top two of EV results. This small compromise in number of patterns was chosen for three main reasons. Firstly, the sample size related to each weather pattern will reduce each time the number of weather patterns increases. This is particularly a problem for the higher numbered weather patterns as they are also the rarest (see Section 4.2). For example, the preferred set of 30 weather patterns (Figure 4) has a sample size of 213 for the highest numbered pattern and 723 for the lowest numbered pattern. This drops to 116 and 655 respectively when increasing the number of patterns to 40. A low sample size means that

weather pattern climatologies can become unreliable. Secondly, having too many patterns means there is a greater likelihood of two or more patterns looking similar. This defeats the purpose of using cluster analysis to split the dominant weather patterns into as set of unique climatological types. It can also make it difficult to objectively identify which weather pattern to assign forecast or analysis fields to when so many weather patterns look the same. Finally, as previously mentioned in this section, the difference in EV scores between 30 and 40 patterns is relatively small and so any benefit in using 40 patterns over 30 is unlikely to be reflected within any forecasting application.

## 4. Results Part 2 – Describing the optimal set of weather patterns

### *4.1. Precipitation climatologies*

The analysis in Section 3 has revealed the preferred set of weather patterns to be produced by clustering *u*- and *v*-component winds at 850-hPa over India Domain 2 (61.5-98.5$^{o}$E and 1.5-37.5$^{o}$N) using a set of 30 weather patterns. These weather patterns, along with their IMD 0.25 degree resolution gridded precipitation climatologies, are shown in Figure 4. Weather pattern precipitation climatologies generated using ERA-Interim at a 1 degree resolution are shown in Figure S1 for comparison and are useful for analysing the spatial precipitation patterns over the sea.

Gridded EV scores (Figure 5) show where the preferred set of weather patterns best represent precipitation variability. Areas with the highest EV scores cover the west coast of India, in a stretch from Mumbai in the north to the coastal side of Nilgiris District in the south. The higher EV scores then stretch across central India, to include Nagpur, and then further east and north to include Darjeeling (in West Bengal state) and Dibrugarh (in Assam state). The high EV scores also stretch

across the Himalayan region, from the Himachal Pradesh and Uttarakhand states in the west to the Sikkim and Assam states in the east. The areas with the highest EV scores correspond well to areas with the highest daily mean precipitation (Figure 6). This suggests that the weather patterns are successfully capturing both the very wet and relatively dry periods in areas where it is wet on average. In contrast, areas where EV scores are low (for example, Jodhpur in North West India, towards the India-Pakistan border, and the area in central southern India, to the east of the Western Ghats) have weather patterns with similar precipitation climatologies. Precipitation here exhibits less extreme seasonal variability with relatively low totals found throughout the year. These low rainfall totals also make these regions less susceptible to hazards associated with extreme precipitation.

*4.2. Historical occurrences*

The preferred set of weather patterns are ordered according to their annual occurrence, with weather pattern 1 occurring most often annually (5.2% of the time) and weather pattern 30 occurring least often annually (1.5% of the time) (Figure 4). The weather patterns exhibit seasonality in their occurrences (Figure 7), with each weather pattern typically occurring for between 3 and 6 months of the year, with near zero occurrences for the other 6 to 9 months. For example, weather patterns 2, 3, 7, 8, 9, 16 and 20 occur in the four months from December through to March and represent the main winter dry period, whereas weather patterns 10, 17, 19 and 21 occur in the four months from June through to September and represent different phases of the active monsoon. However, some patterns have breaks in their months of occurrence. For example, weather patterns 12, 13, 14, 15 and 22 occur for a few months in spring and autumn (during monsoon onset and withdrawal), but not in summer.

*4.3. Persistence and transitions*

Weather patterns persist for between 2 and 3 days on average before transitioning onto another weather pattern (Figure 8). Empirical weather pattern transition probabilities were calculated for all daily lead times between 1 and 9 days. It is not until 2 days onwards that the dominant transitions start to become evident due to patterns normally persisting for at least 2 days. These transitions are often to weather patterns which are in the same broad-scale regime. For example, the weather pattern transition matrix for two days' time (Figure 9) shows that weather pattern 21 typically transitions to weather patterns 17 or 19, which are all variants of the broader-scale active monsoon regime. It is also common to see transitions between two successive broad-scale regimes as the monsoon season evolves throughout the course of the year. For example, one of the most common weather pattern transitions in Figure 9 is from weather pattern 27 (western disturbances) to weather pattern 7 (winter dry period). Weather pattern transition probabilities for longer lead times of around 5 or 6 days onwards (not shown) start to produce a weaker transition signal due to an increasing number of other transitions being possible before these lead times.

### *4.4. Weather regime categories*

Each of the 30 weather patterns has been categorised into one of seven broad-scale regimes (Table 2), called (1) winter dry period, (2) western disturbances, (3) pre/post summer monsoon, (4) monsoon onset, (5) active monsoon, (6) break monsoon and (7) retreating monsoon. Weather patterns were predominantly grouped based on their prevailing months of occurrence, in a similar approach to Islam *et al.* (2018). In this study a prevailing month is defined as one where a weather pattern's occurrence is $\geq$ 5%. In addition, weather pattern wind and precipitation fields were subjectively analysed allowing a further grouping of patterns which have similar spatial characteristics. This combination of

approaches allows for more than one regime category to occur at any given time of the year. For example, the western disturbance and winter dry period regimes both share January, February and March as their prevailing months. Similarly, the monsoon onset, active monsoon and break monsoon regimes all share June and July as their prevailing months. Note that a different definition of prevailing months could lead to a different grouping of weather patterns.

The winter dry period regime is formed from weather patterns 2, 3, 7, 8, 9, 16 and 20, which are most likely to occur between December and March. These patterns all represent very dry conditions across the whole of India. The western disturbances regime is also very dry for most parts, with the exception of the far north. It is formed from weather patterns 5, 23, 24 and 27, which are most likely to occur between January and May. These patterns represent non-monsoonal precipitation events that are responsible for almost one third of the annual precipitation over the northern Indian region and most of the cold season precipitation (Dimri *et al.*, 2015). Western disturbances, which originate over the Mediterranean Sea and the Atlantic Ocean, bring rain or snow to north western parts of India initially, with the risk of heavy precipitation transferring to north-eastern parts during early spring. The spring period, which is still dry for most parts of India, is also often referred to as the hot weather season (Mooley and Shukla, 1987). During these spring months gradual heating of the South Asian land mass takes place and the intertropical convergence zone (ITCZ) moves north, eventually marking the arrival of the summer monsoon in June or July.

The pre/post-summer monsoon regime is formed from weather patterns 12 (mainly pre-monsoon), 13, 14, 15 and 22 (both pre and post-monsoon) which typically occur from May to June for the pre-monsoon period and August to October for the post-monsoon period. This regime represents a mostly dry and hot period just before and after the main summer monsoon, where western disturbances are

mostly absent. Rainfall can be showery in nature and occurs mostly in parts of the southern tip of India, as well as in north-eastern parts where the flow is off the Bay of Bengal. The pre-monsoon period is typically drier for most of the subcontinent with most rainfall restricted to the southern tip of India. This contrasts to the post-monsoon period, where rainfall is related to a weakening of the summer monsoon as winds start to turn northerly across northern India. The flow still remains from the south-west across the Bay of Bengal leading to monsoonal rains continuing along eastern coastal stretches and parts of north-east India.

The monsoon onset regime is formed from weather pattern 26 only and is most likely to occur in June (17.7%) or July (5.6%), with all other months having percentage occurrences < 5%. However, some of the pre-summer monsoon weather patterns described above can also be considered as monsoon onset types if the associated rainfall is persistent enough along the south-west coast of India. The monsoon onset marks the start of the main monsoon season and is associated with a reversal in the direction of the prevailing winds, turning from a dry north-easterly to a moist south-westerly. This allows heavy and persistent rain to start affecting parts of the south-west coast of India, which later progresses across central and northern areas during the active phase. The official onset of the monsoon over India is given as the date when persistent rains arrive over Kerala, which is the state on the far south-west coast of India. The onset over Kerala generally falls on the 1$^{st}$ or 2$^{nd}$ June (Mooley and Shukla, 1987; Rao *et. al.*, 2005), with the onset further north over Mumbai being 10$^{th}$ June (Adamson and Nash, 2013). The monsoon advances north across the subcontinent throughout June and July, resulting in rainfall reaching north-west India by 15$^{th}$ July on average (Tyagi *et. al.*, 2011).

The active monsoon regime is formed from weather patterns 10, 17, 19 and 21 and is most likely to occur from June to September, but is most active during July and August. The average persistence of

these active periods during the monsoon season is about 4 days (Rajeevan *et. al.*, 2010). An active monsoon period sees the potential for rainfall to cover many central and north-eastern areas, as well as down most of the west coast. Mumbai (on the west coast) experiences some of its highest rainfall totals during the active monsoon, with the rainfall distribution for all weather patterns at Mumbai shown in Figure 10. Here it is evident that the four weather patterns with the highest daily mean rainfall are active monsoon types. It is encouraging to see that daily mean rainfall totals vary between these four types, with weather pattern 19 being the wettest (with a mean of 46.4 mm; Table S1) and weather pattern 10 having the lowest rainfall (with a mean of 22.9 mm; Table S1). Rainfall variability between the active monsoon weather patterns allows identification of periods with a higher risk of flooding impacts.

The break monsoon regime is formed from weather patterns 4 and 11 and is most likely to occur from June to August. During the main summer monsoon season there are periods when the monsoon trough shifts closer to the foothills of the Himalayas, therefore interrupting the moist south-westerly monsoon flow. This causes winds to become more westerly or north-westerly across northern India leading to a sharp decrease in rainfall over most parts of the country. However, rainfall increases along the foothills of the Himalayas, Northeast India and parts of the Southern Peninsula (e.g. parts of Rayalaseema and Tamil Nadu). Rajeevan *et al.* (2010) found that break spells have an average persistence of 6 days. Blanford (1886) first identified these break periods as 'intervals of droughts'. Since then, these periods have been called 'breaks' by Indian meteorologists for over a hundred years (e.g., Raghavan (1973); Krishnamurti and Bhalme (1976); Sikka (1980), Gadgil and Joseph (2003)).

The final regime to be identified is called the retreating monsoon, which is formed from weather patterns 1, 6, 18, 25, 28, 29 and 30. These patterns typically occur from September through to

December. The withdrawal of the summer monsoon typically begins around the second week in September (Rao *et al.*, 2005). By the beginning of October a weak area of high pressure normally forms over the Tibetan Plateau which pushes dry air south towards India. Simultaneously, the ITCZ moves south allowing high pressure to form over northern India by mid-October. This causes the upper level winds over Northern India to turn from westerly to easterly ushering in the start of dry north-easterly winds at the surface. As these north-easterly winds flow over the Bay of Bengal they pick up moisture, which falls as monsoon rains over the southern tip of India. These retreating monsoon rains normally begin in coastal Tamil Nadu during mid-to-late October, and later go on to affect most southern states of India, which often receive most of their rainfall from the retreating monsoon. Chennai, which is on the south-east coast of India and one of the 10 locations in Table 1, gets most of its rainfall during the retreating monsoon. Figure 11 shows the rainfall distribution for each weather pattern at Chennai, where most rainfall comes from weather pattern 18, with a daily mean rainfall of 22.5 mm (Table S1). This weather pattern is one of the several retreating monsoon patterns, with many of the corresponding patterns (1, 6, 25, 28, 29 and 30) experiencing daily mean rainfall totals < 10 mm. This suggests that these weather patterns are able to distinguish between the wettest and relatively drier phases of the retreating monsoon at a given location.

## 5. Discussion

We propose a set of 30 weather patterns (Figure 4) based on 850-hPa winds (*u*- and *v*-components) with a spatial domain covering the entire Indian sub-continent (India Domain 2; Figure 1). This weather pattern set, as with all others tested, was generated by clustering daily data covering the full year. It would have also been possible to apply the clustering to days in separate months or seasons, such as generating separate weather patterns for each two-month period throughout the year, as done

by Fereday *et al.* (2008) for Europe. However, this approach is not suitable for this study due to complications it would cause with future weather pattern forecasting applications, whereby forecasting weather pattern transitions between two periods would become problematic. Results show that the preferred set of weather patterns exhibit significant seasonality in their occurrences anyway, with a collection of half a dozen or more being much more likely than others to occur at any given time of year. The approach in this study also gives the opportunity for any of the weather patterns to occur at any time during the year. This could be useful when identifying seasonal changes in the occurrence of a given weather pattern, as might be caused by climate change.

EV results (Table 1) show that weather patterns generated using *u*- and *v*-component winds provide the best representation of precipitation variability, which reflects how much India's climate is influenced by changes in the direction of the prevailing wind. Clustering on three wind levels (10-m, 925-hPa and 850-hPa) was tested, with results showing that weather patterns generated using winds at higher elevations perform better in terms of representing precipitation variability. However, weather patterns generated using winds at 850-hPa are only marginally better than those generated using winds at 925-hPa (Table 1), suggesting that any benefit in increasing the wind elevation further is likely to be minimal. Winds at the 850-hPa pressure level are about 1.5 km above sea-level meaning that the representation of winds over mountainous regions such as the Western Ghats and Himalayan region are likely to be better. The effect of surface friction on wind speed is also dramatically reduced at this level. For example, wind speed at lower elevations, particularly 10-m, will experience a slow down over land due to friction. This could influence the resultant weather patterns by biasing the large-scale circulation patterns towards wind characteristics over the sea where the effects of friction are less. Wind components at 850-hPa were also used by Roller and Qian (2016) to generate a set of

five weather patterns over the Northeast United States, where the focus was on wintertime circulation patterns and their relationship with storm tracks and precipitation variability.

Weather pattern studies for Europe, e.g. Beck *et al.* (2013) and Neal *et al.* (2016), have shown that weather patterns generated over smaller areas tend to relate better with precipitation variability. This is because precipitation over Europe is closely related to the circulation at a relatively small scale, in contrast to temperature which is more closely related to the circulation at a relatively large-scale. However, results in this study show that the same is not always true for India, at least for precipitation variability which was tested here. For example, EV results in Figure 3 show that precipitation variability at Darjeeling is best represented by weather patterns using large domains defined over the whole of India, closely followed by domains centred over Darjeeling itself as well as Nilgiris in southern India, suggesting a close relationship between weather patterns defined over southern India and precipitation variability in northern India. In contrast, EV results in Figure 2 show that precipitation variability at Nilgiris is best represented by weather patterns defined over a relatively small area centred on Nilgiris itself, followed by the India domains in second place and the Darjeeling domains far behind in third place. These results suggest that the driving mechanisms behind precipitation variability over India are defined by varying synoptic scales depending on location. The reason for the good south to north relationship may be due to the 40-day south-to-north propagation of rainfall during the active and break phases of the summer monsoon (Yasunari, 1981), which means that rainfall in the south is strongly correlated with rainfall in the north. For example, during the active monsoon (e.g. weather patterns 19 and 21; Figure 4) precipitation maxima occur along the west coast and across central and eastern areas, but it is relatively dry in the north. In contrast, during the break monsoon (e.g. weather patterns 4 and 11; Figure 4) the heaviest rainfall has stopped in the south and moved north towards the Himalayan region. Conversely, in northern India westerly propagating

monsoon depressions are responsible for significant rainfall and can contribute up to 60% of total summer monsoon rainfall (Praveen *et al.*, 2015). These systems have more local drivers, such as the mean state monsoon circulation within the monsoon trough region **(**Levine and Martin, 2018) and Himalayan orography (Hunt and Parker, 2016) that are less connected with southern India. Therefore, this may explain why weather patterns defined over north east India have a weak relationship with rainfall variability in southern India.

The decision on the best overall weather pattern domain to use is based on interpretation of all EV scores (Figures 2 and 3), where differences in scores between the preferred and runner-up domains are normally relatively small. For example, precipitation variability at Darjeeling using a set of 30 patterns defined over the whole of India have an EV score of around 0.35, compared to around 0.3 for smaller domains centred on Darjeeling (Figure 3). Similarly for Nilgiris, the domains defined over the whole of India have an EV score of around 0.27, compared to around 0.35 for smaller domains centred on Nilgiris (Figure 2). This suggests that using one large weather pattern domain to represent precipitation variability at multiple locations across India is likely to be a good compromise for those locations where smaller domains perform better. Using a large domain also increases the opportunity to use one set of weather patterns for as many applications as possible across the whole of India.

The main purpose for generating the preferred set of weather patterns presented here was to improve the understanding of precipitation variability over India. The results of this study mean it is now possible to objectively identify which weather patterns within the active monsoon regime bring the most rainfall to any given location. The same also applies for the other phases of the Indian climate such as the retreating monsoon and western disturbances. In total, seven weather regime categories were identified (Table 2), with many comparable to the eight groups produced by Islam *et al.* (2018)

which are described as (1) the winter period, (2) the pre-monsoon period, (3) the monsoon period and (4) the post-monsoon period, as well as the four transitions between these groups. All weather regime categories produced in this study have more than one weather pattern with the exception of the monsoon onset which is formed from weather pattern 26 only. However, some of the pre-monsoon types which occur during May and June could also be considered monsoon onset types in some areas and the way the weather patterns were categorised to some extent was a subjective process.

As well as relating to precipitation variability, these weather patterns are intended for use in future research by relating them to the occurrence of high impact weather and meteorologically induced hazards such as landslides. The weather patterns are also suitable for use within probabilistic medium-to long-range weather pattern forecasting tools driven by global ensemble prediction systems, such as those run operationally by many national forecasting agencies. For example, in 2018 the National Centre for Medium-Range Weather Forecasts (NCMRWF) in India operationally implemented a 12 km resolution global ensemble prediction system with 23 ensemble members and a 10 day forecast lead time (Mamgain *et al.*, 2018 and Kumar *et al.*, 2018). Here, multiple forecast scenarios (ensemble members) for each daily forecast lead time or collective forecast period can be objectively assigned to the closest matching weather pattern definition, allowing forecast probabilities for each weather pattern to be derived. Similar forecasting tools are already available operationally over Europe as described in the Introduction.

The weather regime or pattern definitions used in the European applications are based on single pressure fields, which allows an easy comparison with the forecast pressure fields from ensemble members when identifying the closest match. This forecast-pattern assignment can be achieved using methods such as spatial correlation (as used by Ferranti and Corti, 2011) or a measure of the intensity

difference in features (areas of low and high pressure) between forecast fields and the regime or pattern definition fields (sometimes referred to as 'distance'; as used by Neal *et al.*, 2016). In this study, weather patterns are defined using their wind *u*- and *v*-components. These two fields could be converted into one single field by deriving their stream functions, which could be done using the Windspharm Python package created by Dawson (2016). The stream functions would also need to be calculated for the forecast fields, which would then allow similar assignment methods to be used as in the European forecasting products. The seven weather regime categories (Table 2) can also be used in any future forecasting product by aggregating up probabilities from each of the 30 patterns depending on which of the seven categories they are mapped to. This would be useful for identifying general forecast trends towards a particular broad-scale regime when ensemble spread among the 30 patterns is high; something potentially more useful at longer forecast lead times.

## 6. Conclusion

The aim of this study was to derive an optimal set of weather patterns for representing precipitation variability over India, by applying k-means clustering to daily reanalysis fields from ERA-Interim between 1979 and 2016. The optimal set of weather patterns was arrived at by testing 192 sets of weather patterns against their ability to represent precipitation variability at 10 geographically varying locations across India using gridded precipitation analysis data from IMD. The different pattern sets were arrived at by varying the parameter clustered (PMSL and *u*- and *v*-component winds at three vertical levels), the domain clustered (12 sizes in total) and finally the number of resultant patterns (10, 20, 30 and 40). The final optimal set of weather patterns was a set of 30 (Figure 4), which was arrived at by clustering *u*- and *v*-component winds at the 850-hPa pressure level, using an India-wide domain covering the area from 61.5-98.5$^{o}$E and 1.5-37.5$^{o}$N.

The optimal set of weather patterns persist for between two and three days before transitioning to another pattern (Figure 3), which can be a transition to a similar pattern or a transition to a completely new regime (Figure 4). The weather patterns can be attributed to one of seven broad-scale circulation categories (regimes) including the winter dry period, western disturbances, the pre- and post-summer monsoon, the monsoon onset, the active monsoon, the break monsoon and the retreating monsoon. All categories have a collection of weather patterns, with the exception of the monsoon onset, which is formed from weather pattern 26 only. Identifying weather pattern variants within most of the regime categories is important for establishing the specific circulation types behind the largest rainfall totals and resulting hazards such as flooding or landslides (e.g. identifying the most flood prone periods within the active monsoon regime). The weather patterns may also be useful for relating non-meteorological data such as from the agriculture, transport or energy sectors, to help understand how large-scale atmospheric circulation affects specific industries.

The weather patterns represent precipitation variability best in areas which get the most precipitation (Figures 4 and 5). For example, the rainfall distribution among weather patterns is most different along the west coast (in an area running from Mumbai in the north to Nilgiris in the south). This high variability in precipitation distribution among weather patterns extends into central and eastern areas of India, as well as up towards the Himalayan region in the far north (including areas such as Uttarakhand and Darjeeling which have high sensitivity to landslides). Areas with low precipitation variability among weather patterns includes the dry regions in northeast India and areas sheltered by the Western Ghats in central-southern India.

The clustering produced daily historical weather pattern classifications from 1979 to 2016, which can be used to reproduce the weather pattern definition fields by taking a mean of all daily ERA-Interim reanalysis fields assigned to each weather pattern. The historical classifications can then be extended to present by designing a method to automatically assign reanalysis fields after 2016 to the closest matching weather pattern definition – methods introduced for a forecast product in Section 5 could be used here. This would then open up the weather patterns for use in further research by relating them to the occurrence of past high impact weather and meteorologically induced hazards such as landslides. The weather patterns are also suited for use within probabilistic medium- to long-range weather pattern forecasting tools driven by ensemble prediction systems, whereby forecast scenarios (ensemble members) can be objectively assigned to the closest matching weather pattern definition. Once weather pattern characteristics are understood in terms of their meteorological climatologies or sector-specific impacts (e.g. impacts on agriculture, energy or transport), it then becomes easier to interpret forecast output and describe likely consequences. The number of forecasting applications is potentially large. Creation of a forecasting product would also enable objective analysis into the predictability of different Indian weather patterns or regime groupings.

Future research could also consider experimenting with other clustering methods, such as SOMs, and then using a metric the same as or similar to the EV score used here to see how representative the resultant patterns are as far as precipitation variability is concerned. Similarly, other parameters (or even combinations of parameters) could be tested within the clustering methodology to see if they also produce representative resultant patterns. For example, it is possible that a humidity variable representing atmospheric moisture capacity may also relate well to precipitation variability. However, it is likely that moisture information is implicitly included in the weather patterns presented here because it is so directly related to wind direction.

## 7. Data download

The weather pattern definitions (as mean *u*- and *v*-component wind composites at 850-hPa) for the optimal set of weather patterns and their associated daily historical classifications between 1979 and 2016 are available for download from the PANGAEA online data repository (Neal *et al.*, 2019).

## 8. Acknowledgments

This work was carried out as part of the LANDSLIP (Landslide multi-hazard risk assessment, preparedness and early warning in South Asia integrating meteorology, landscape and society) project which had two UK grant funders: NERC and DFID (Grant numbers: NE/P000681/1 and NE/P000649/1). This work and some of its contributors (Robert Neal, Joanne Robbins and Rutger Dankers) were also supported by the Met Office Weather and Climate Science for Service Partnership (WCSSP) India Programme as part of the Newton-Bhabha Fund. The authors would like to thank the two reviewers for their helpful comments on an earlier version of this paper. Thanks also go to Gill Martin for useful discussions around the relationship between weather patterns defined over southern India and precipitation variability in northern India.

## 9. References

Adamson GCD and Nash DJ. 2013. Long-term variability in the date of monsoon onset over western India. *Climate Dynamics*. 40: 2589-2603

Andenberg MR. 1973. Cluster Analysis for Applications. A volume in Probability and Mathematical Statistics: A Series of Monographs and Textbooks. *Academic Press, New York.*

Annamalai H, Slingo JM, Sperber KR, Hodges K. 1999. The mean evolution and variability of the Asian summer monsoon: comparison of ECMWF and NCEP-NCAR reanalysis. Monthly Weather Review. 127: 1157-1186.

Bao M, Wallace J. 2015. Cluster Analysis of Northern hemisphere Wintertime 500-hPa Flow Regimes during 1920-2014. *Journal of the Atmospheric Sciences*. 72: 3597-3608

Beck C and Philipp A. 2010. Evaluation and comparison of circulation type classifications for the European domain. *Physics and Chemistry of the Earth*. 35: 374-387

Beck C, Philipp A, Streicher F. 2013. The effect of domain size on the relationship between circulation type classifications and surface climate. *Int. J. Climatol.* 15: 3687–3703.

Bedi HS, Bindra MMS. 1980. Principal components of monsoon rainfall. *Tellus*. 32: 296 - 298.

Blanford HF. 1886. Rainfall of India. Mem. Ind. Met. Dept. 2: 217-448

Casado MJ and Pastor MA. 2013. Circulation types and winter precipitation in Spain. *International Journal of Climatology*. 36: 2727-2742

Chattopadhyay R, Sahai AK, Goswami BN. 2008. Objective Identification of Nonlinear Convectively Coupled Phases of Monsoon Intraseasonal Oscillation: Implications for Prediction. *Journal of the Atmospheric Sciences*. 65: 1549-1569

Dawson A. 2016. Windspharm: A High-Level Library for Global Wind Field Computations Using Spherical Harmonics. Journal of Open Research Software, 4(1), p.e31. DOI: http://doi.org/10.5334/jors.129

Dee DP, Uppala S, Simmons A, Berrisford P, Poli P, Kobayashi S, *et al.* 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc*. 137: 553-597

Dimri AP, Niyogi D, Barros AP, Ridley J, Mohanty UC, Yasunari T, Sikka DR. 2015. Western Disturbances: A review. *Rev. Geophys*. 53: 225-246

Dorling SR, Davies TD, Pierce CE. 1992. Cluster analysis: A technique for estimating the synoptic meteorological controls on air and precipitation chemistry—Method and applications. *Atmospheric Environment*. 26: 2574-2581

Ferranti L, Corti S. 2011. New clustering products. *ECMWF Newsl*. 127:6-11

Fereday DR, Knight JR, Scaife AA, Folland CK, Philipp A. 2008. Cluster analysis of North Atlantic–European circulation types and links with tropical Pacific sea surface temperatures. J. Clim. 21: 3687–3703.

Ferranti L, Corti S, Janousek M. 2015. Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Q. J. R. Meteorol. Soc*. 141:916–924, DOI: 10.1002/qj.2411

Gadgil S and Gadgil S. 2006. The Indian Monsoon, GDP and Agriculture. *Economic and Political Weekly*. 41: pp. 4887-4895

Gadgil S and Joseph PV. 2003. On breaks of the Indian monsoon. Journal of Earth System Science. 112:529-558 DOI: 10.1007/BF02709778

Gadgil S, Yadumani, Joshi NV. 1993. Coherent rainfall zones of the Indian region. *International Journal of Climatology*. 13: 547-566.

Hunt KMR, Parker DJ. 2016. The movement of Indian monsoon depressions by interaction with image vortices near the Himalayan wall. *Quarterly Journal of the Royal Meteorological Society*. 142: 2224-2229

Islam MR, Sheridan SC, Lee CC. 2018. Using self-organising maps to identify the South Asian seasonal cycle. *Theor Appl Climatol*. pp1-17

Jones E, Oliphant E, Peterson P, *et al.* 2001. SciPy: Open Source Scientific Tools for Python, http://www.scipy.org/ [Online; accessed 2018-06-11].

Kalnay et al. 1996. The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc*. 77: 437-471

Krishnamurti TN and Bhalme HN. 1976. Oscillations of a monsoon system. Part 1. Observational aspects. *J. Atmos*. Sci. 33: 1937-1954

Krishnamurthy V, Shukla J. 2000. Intraseasonal and interannual variability of rainfall over India. *Journal of Climate*. 13: 4366-4377.

Kumar S, Jayakumar A, Bushair MT, Buddhi Prakash J, George G, Lodh A, Indira Rani S, Mohandas S, George JP, Rajagopal EN. 2018. Implementation of New High Resolution NCUM Analysis-Forecast System in Mihir HPCS. NCMRWF Tech. Report No. NMRF/TR/01/2018. Available from NCMRWF at www.ncmrwf.gov.in. Accessed 5th March 2019.

Levine RC and Martin GM. 2018. On the climate model simulation of Indian monsoon low pressure systems and the effect of remote disturbances and systematic biases. *Climate Dynamics*. 50: 4721-4723

Mamgain A, Sarkar A, Dube A, Arulalan T, Chakraborty P, George JP, Rajagopal EN. 2018. Implementation of Very High Resolution (12 km) Global Ensemble Prediction System at NCMRWF and its Initial Validation. NCMRWF Tech. Report No. NMRF/TR/02/2018. Available from NCMRWF at www.ncmrwf.gov.in. Accessed 5th March 2019.

Martha TR, Roy P, Govindharaj KB, Kumar KV, Diwakar PG, Dadhwal VK. 2015. Landslides triggered by the June 2013 extreme rainfall event in parts of Uttarakhand state, India. *Landslides*. 12: 135-146

Mishra V, Shah H. 2018. Hydroclimatological Perspective of the Kerala Flood of 2018. *Journal of the Geological Society of India*. 92: 645-650

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. R. Meteorol. Soc*. 122: 73-119

Mooley DA, Shukla J. 1987. Variability and forecasting of the summer monsoon rainfall over India. In: Chang C-P, Krishnamurti TN (eds). Monsoon meteorology. *Clarendon Press, Oxford*, pp 26–59

Neal R, Fereday D, Crocker R, Comer R. 2016. A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*. 23: 389-400

Neal R, Dankers R, Saulter A, Lane A, Millard J, Robbins G, Price D. 2018. Use of probabilistic medium- to long-range weather pattern forecasts for identifying periods with an increased likelihood of coastal flooding around the UK. *Meteorological Applications.* DOI:10.1002/met.1719.

Neal R, Robbins J, Dankers R, Mitra A, Jayakumar A, Rajagopal EN, Adamson George. 2019. Weather pattern definitions for India and their daily historical classifications (1979 to 2016). PANGAEA, https://doi.org/10.1594/PANGAEA.902030

Pai DS, Sridhar Latha, Rajeevan M, Sreejith OP, Satbhai NS, Mukhopadhyay B. 2014. Development of a new high spatial resolution (0.25° × 0.25°) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *Mausam*. 65: 1-18

Praveen V, Sandeep S, Ajayamohan RS. 2015. On the relationship between mean monsoon precipitation and low pressure systems in climate simulation models. *Journal of Climate*. 28: 5305-5324

Prein AF, Holland GJ, Rasmussen RM, Clark MP, Tye MR. 2016. Running dry: The U.S. Southwest's drift into a drier climate state. *Geophysical Research Letters*. 43: 1272-1279

Raghavan K. 1973. Break-Monsoon over India. *Mon. Wea. Rev*. 101: 33-43

Rajeevan M, Gadgil S, Bhate J. 2010. Active and break spells of the Indian summer monsoon. *Journal of Earth System Science*. 119: 229–247

Rao PLS, Mohanty UC, Ramesh KJ. 2005. The evolution and retreat features of the summer monsoon over India. *Meteorol Appl.* 12: 241–255

Roller CD, Qian J. 2016. Winter Weather Regimes in the Northeast United States. *J Clim*. 29:2963–2980

Sikka DR. 1980. Some aspects of the large scale fluctuations of summer monsoon rainfall over India in relation to fluctuations in the planetary and regional scale circulation parameters. *Proc. Indian Acad. Sci. (Earth Planet. Sci.)*. 89: 179-195

Tyagi A, Mazumdar AB, Khole M, Gaonkar SB, Devi S. 2011. Redetermination of normal dates of onset of southwest monsoon over India. *Mausam.* 62: 321–328

Yasunari T. 1981. Structure of an Indian summer monsoon system with around 40-day period. *J Meteorol Soc Japan*. 59: 336-354

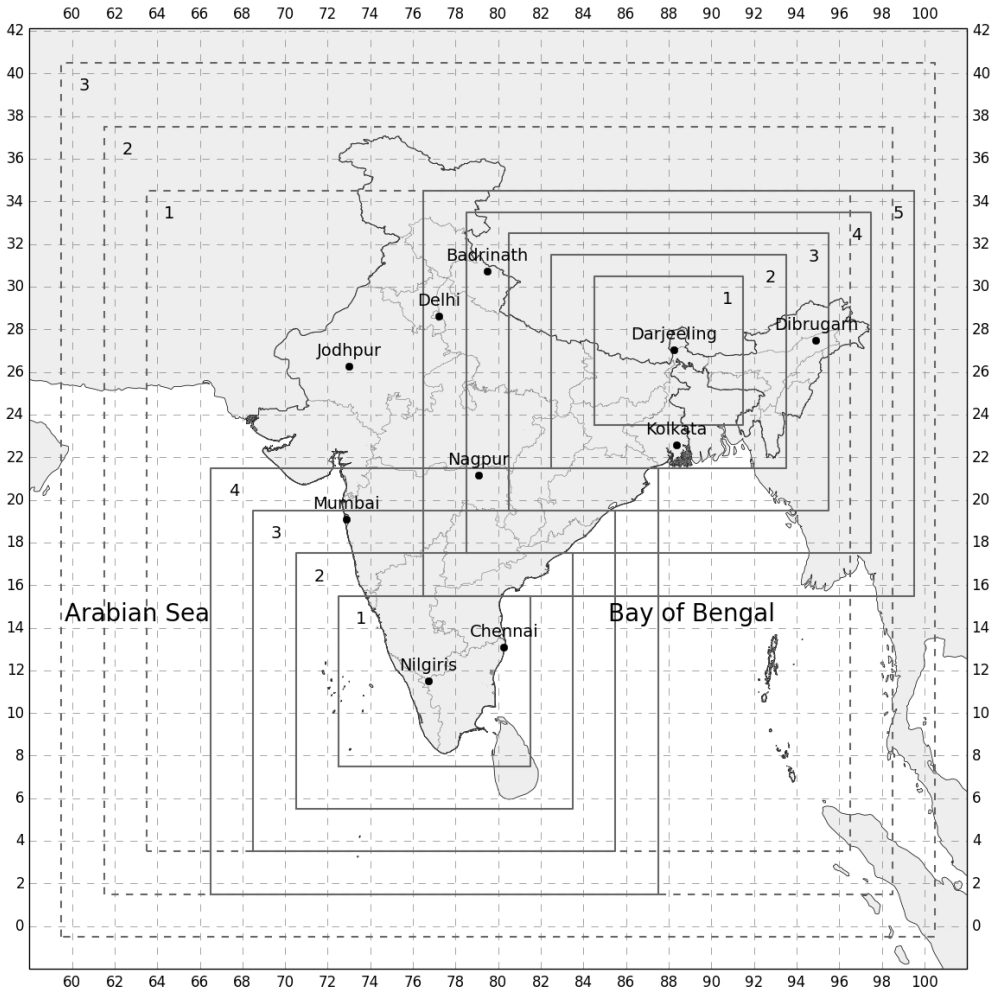


**Figure 1:** Weather pattern domains used in the clustering experiments. Weather pattern boundaries mark the edge of the 1 × 1° resolution ERA-Interim grid.

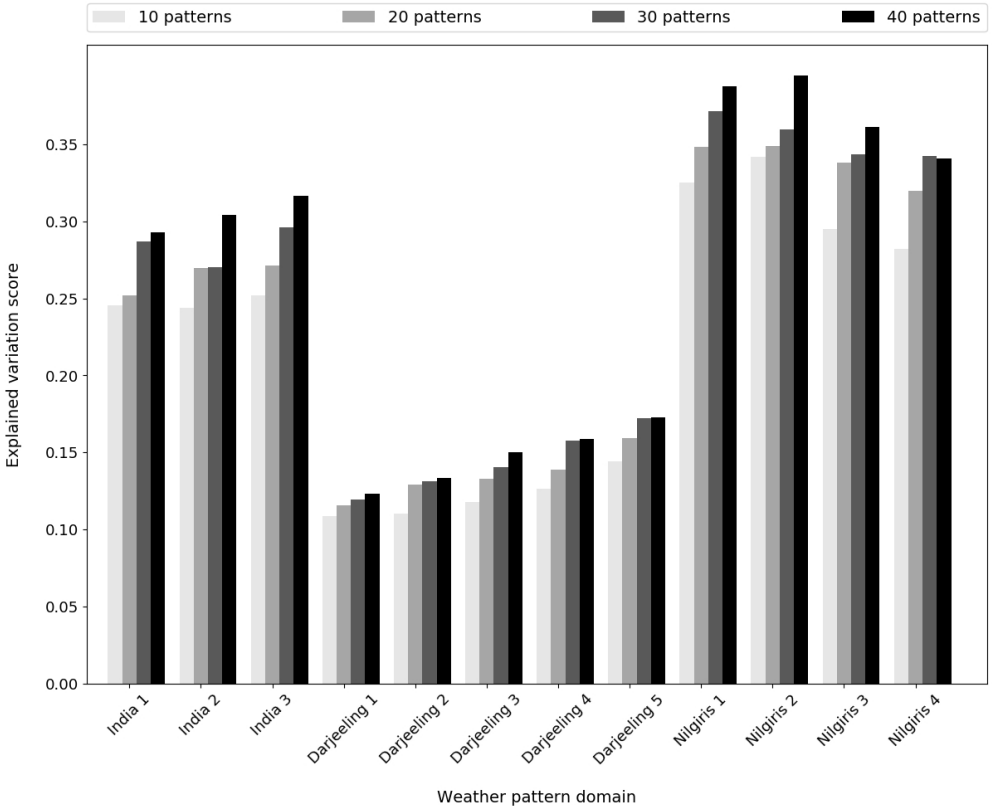


**Figure 2:** Explained variation showing the ability of each weather pattern set (defined using *u* and *v*-component winds at 850-hPa) to represent precipitation variability at Nilgiris using IMD's 0.25 degree resolution gridded rainfall observation data set between 1979 and 2016. For an overview of the domains see Fig. 1. Daily rainfall fields were neighbourhood post-processed using a two grid cell square neighbourhood before deriving the explained variation.

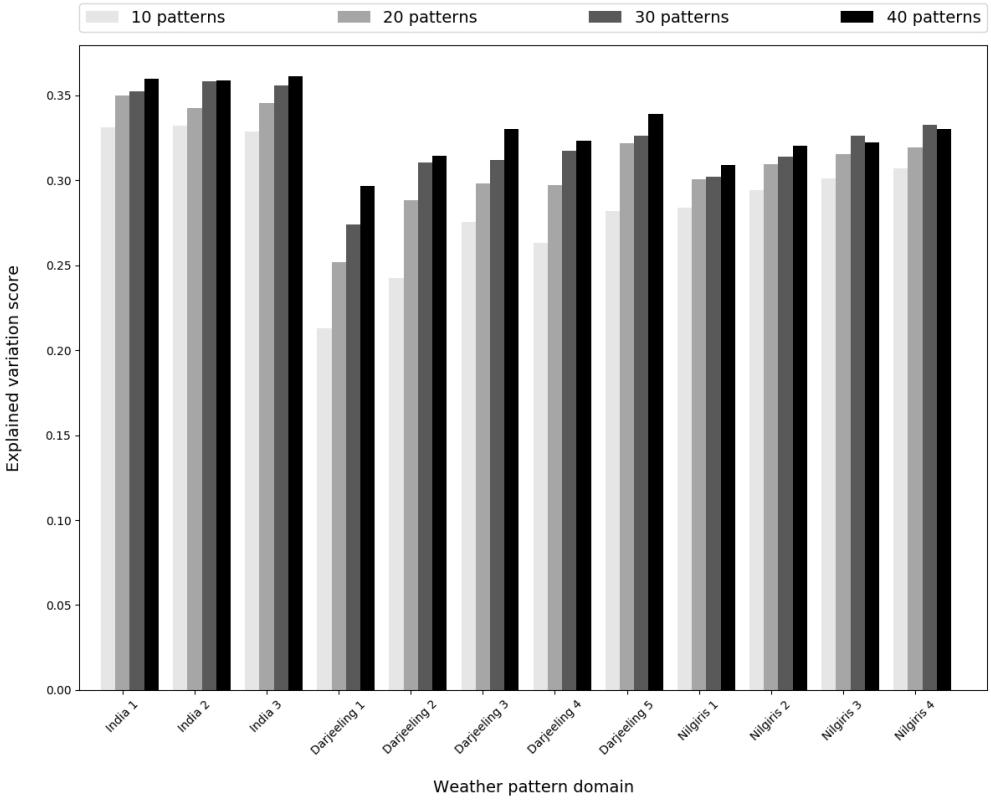


**Figure 3:** As in Fig. 2 but for Darjeeling.

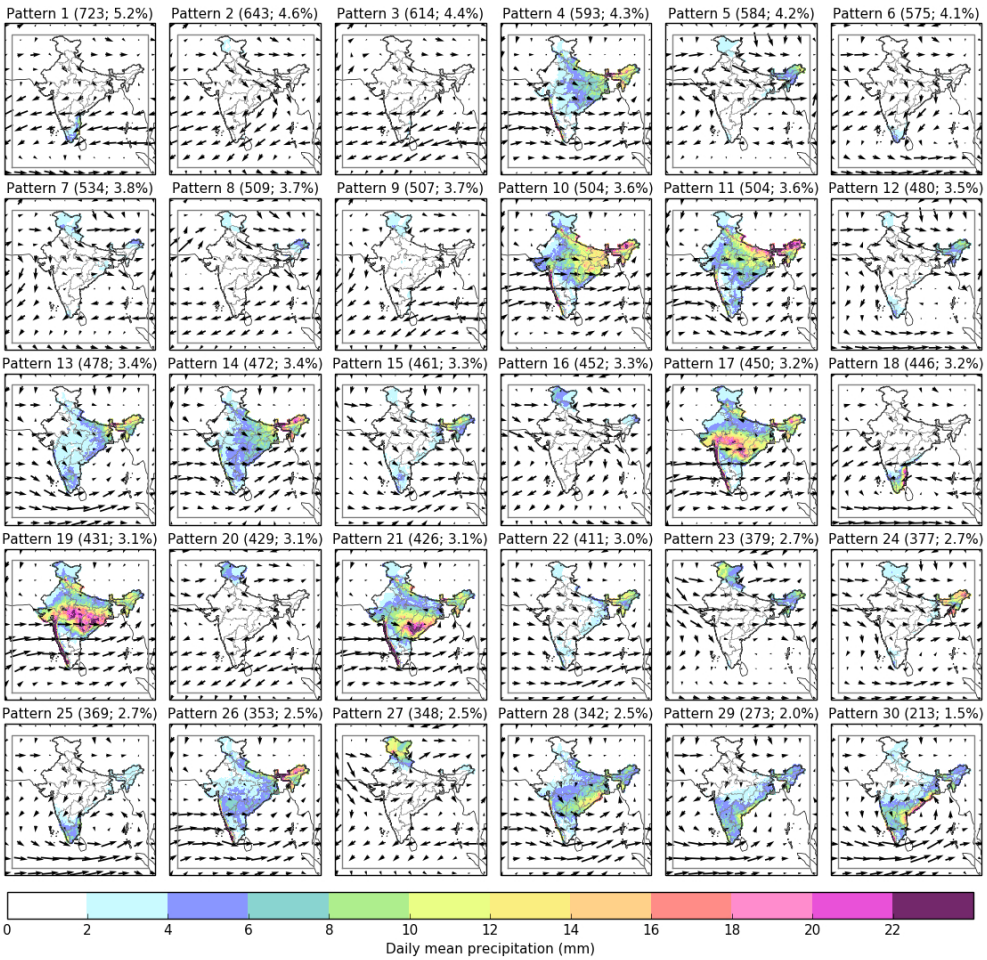


**Figure 4:** Weather pattern definition maps for the preferred set of weather patterns. Arrows represent wind speed and direction at 850-hPa from ERA-Interim. Coloured contours show daily mean rainfall from IMD's 0.25° resolution gridded observation dataset–data is plotted at grid resolution with no neighbourhood post-processing. The inner box shows the area used in the clustering. Numbers in brackets give the sample size followed by mean annual occurrence for each weather pattern. All data is valid for the period 1979 to 2016.

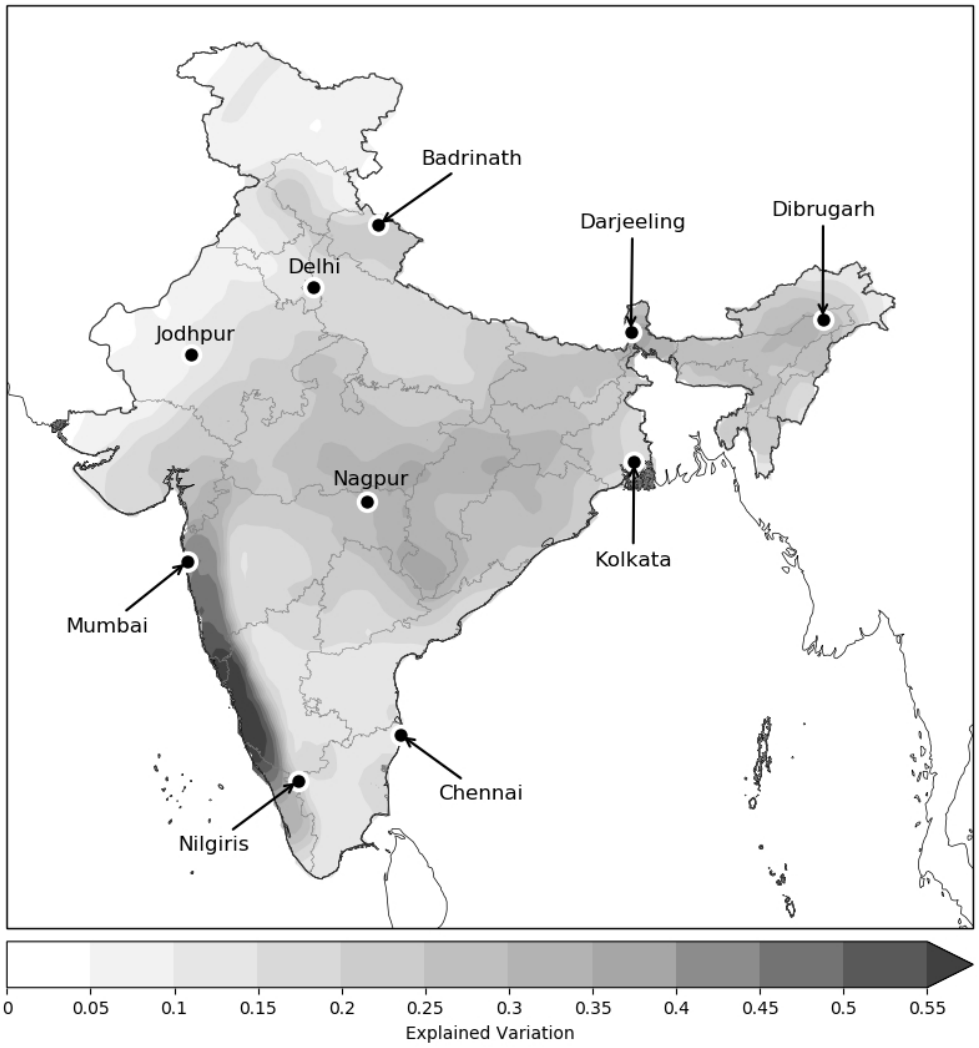


**Figure 5:** Explained variation for the preferred set of weather patterns. Explained variation is based on the distribution of rainfall across weather patterns using IMD's 0.25° resolution gridded rainfall observation data set between 1979 and 2016. Daily rainfall fields were neighbourhood post-processed using a two grid cell square neighbourhood before deriving the explained variation.

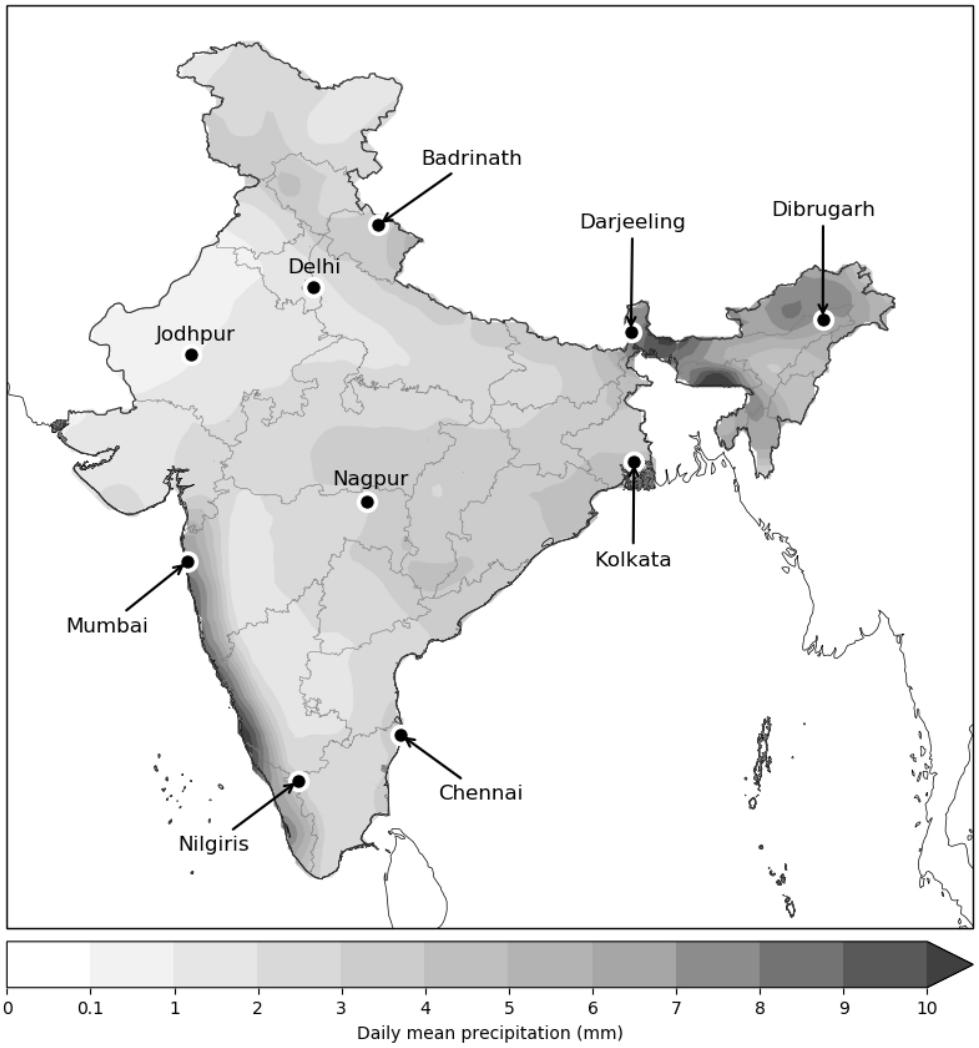


**Figure 6:** Daily mean precipitation using IMD's 0.25° resolution gridded rainfall observation data set between 1979 and 2016. Grid point values represent a mean of all points within a two grid cell square neighbourhood.

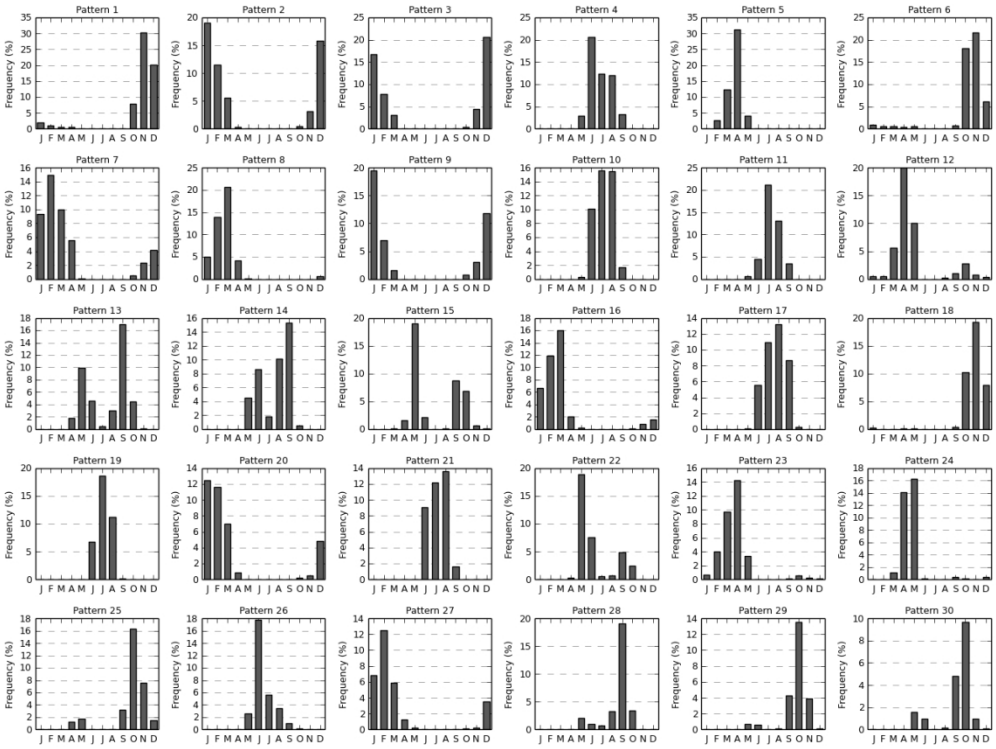


**Figure 7:** Monthly occurrences for each weather patterns (from the preferred set of weather patterns) for the period 1979 to 2016.

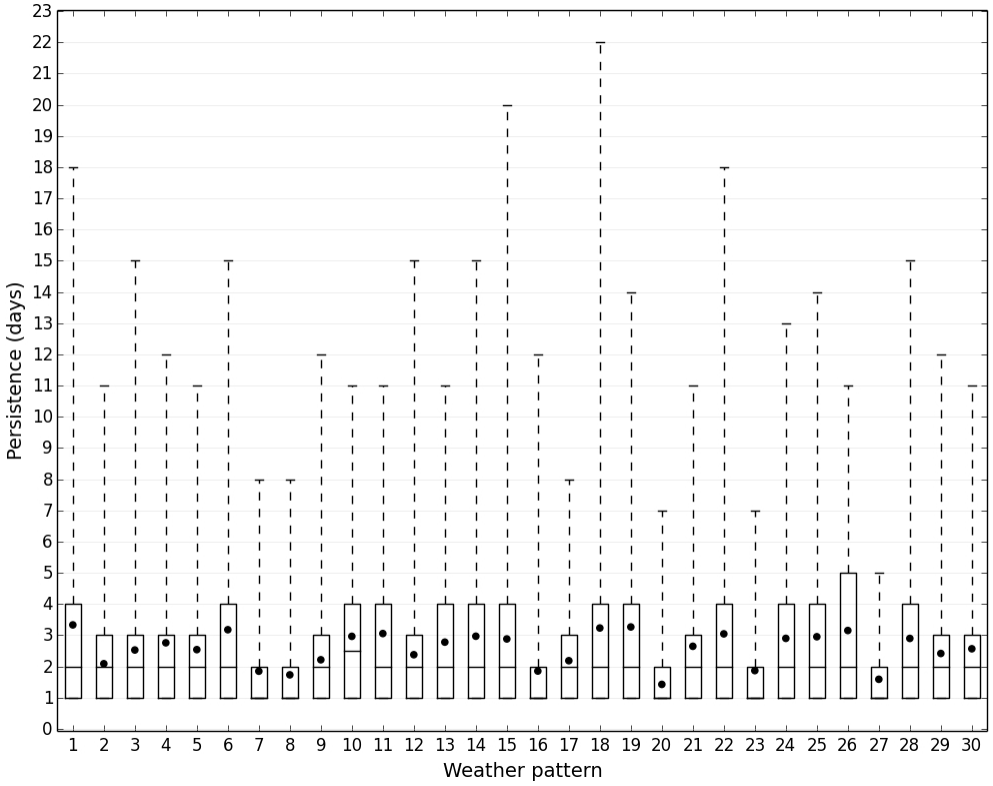


**Figure 8:** The mean persistence for each weather pattern (using the preferred set of weather patterns) for the period 1979 to 2016. Boxes give the 25th, 50th and 75th percentiles. Whiskers give the historical maximum and minimum persistence. Dots give the mean persistence.

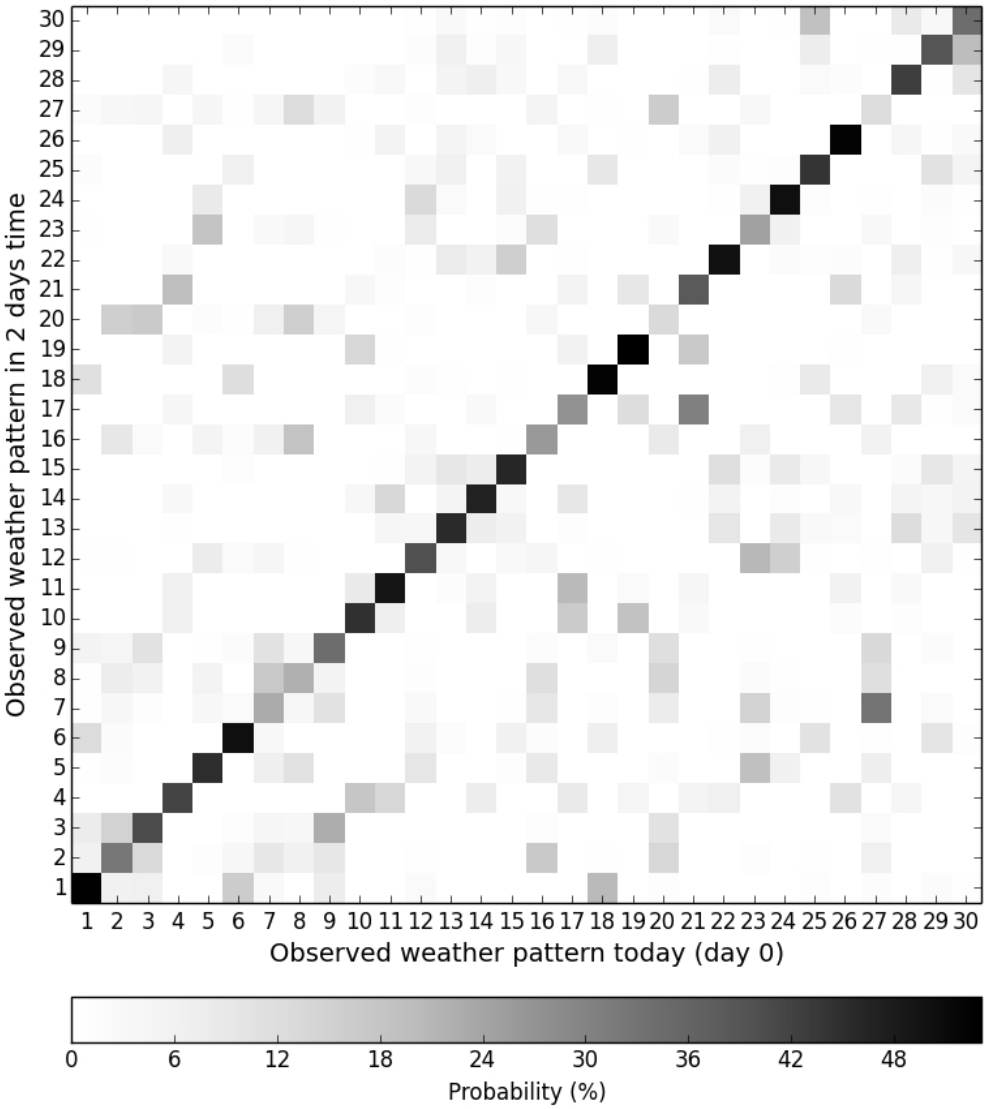


**Figure 9:** A two day weather pattern transition matrix (using the preferred set of weather patterns) for the period 1979 to 2016.

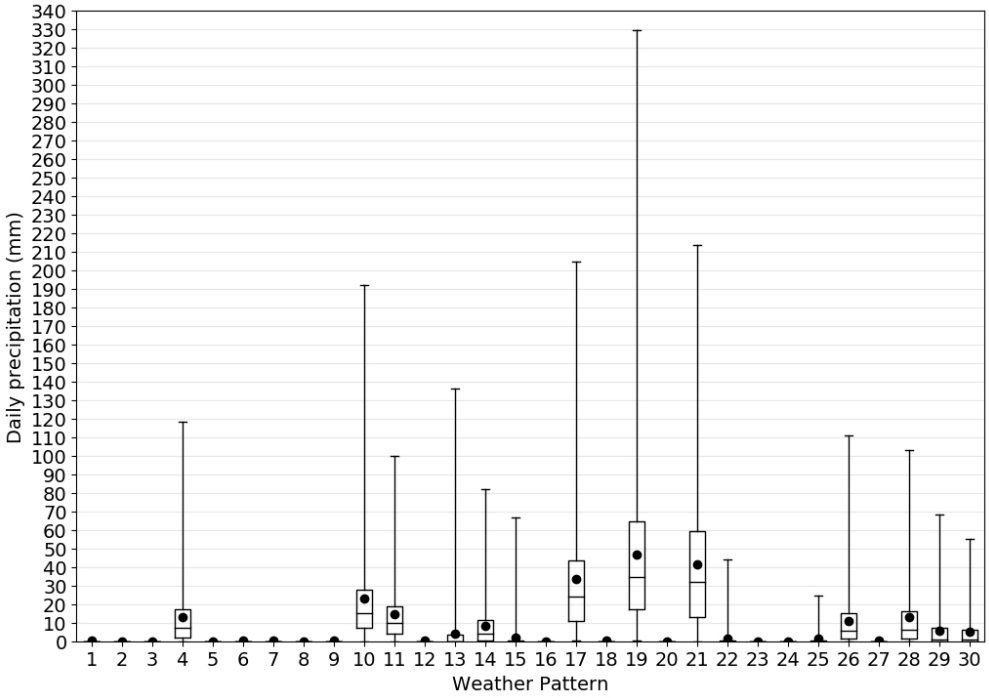


**Figure 10:** Box and whisker plots showing the distribution of rainfall among the preferred set of weather patterns at Mumbai using IMD's 0.25° resolution gridded rainfall observation data set for the period 1979 to 2016. Data values for Mumbai represent a mean of all points within a two grid cell square neighbourhood. Boxes provide the 25th, 50th and 75th percentiles. Whiskers provide the maximum and minimum values. Dots provide the mean.

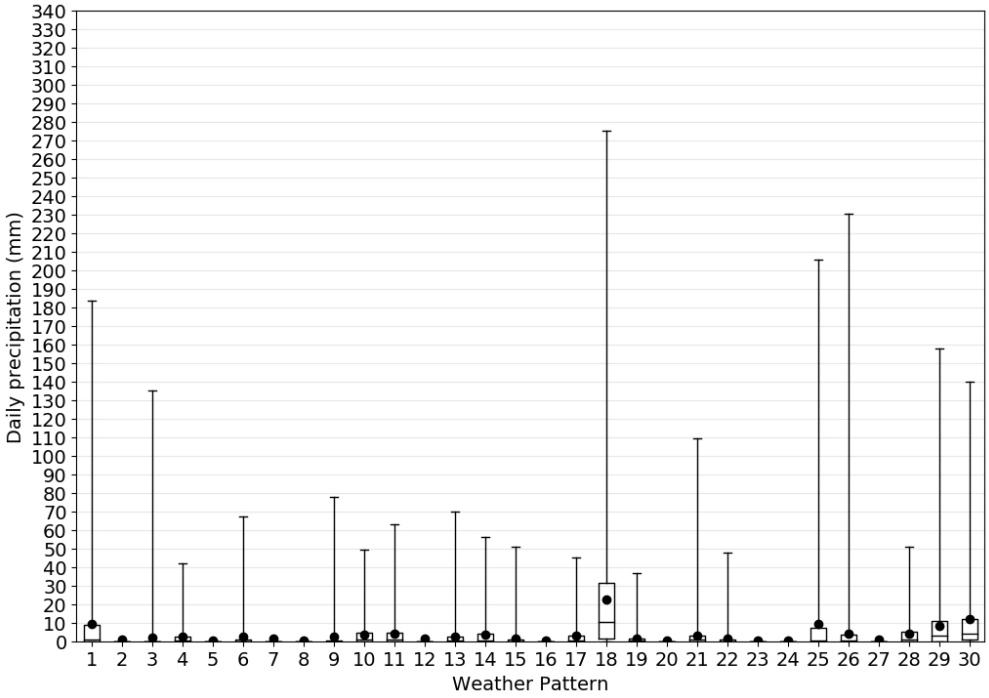


**Figure 11:** As in Fig. 10 but for Chennai.

**Table 1:** Ranking of clustering parameters for a selection of sites across India assuming a set of 30 weather patterns are used. The rankings are based on the parameter with the highest explained variation using IMD's 0.25 degree resolution gridded rainfall observation data set for the period 1979 to 2016. The parameter with the overall lowest score across all sites has the best explained variation.

| | Latitude and longitude | Elevation (m) | 10 m wind ranking | 925 hPa wind ranking | 850 hPa wind ranking | PMSL ranking |
|---|---|---|---|---|---|---|
| **Darjeeling** | 27.04, 88.26 | 1952 | 3 | 4 | 1 | 2 |
| **Dibrugarh** | 27.47, 94.91 | 109 | 2 | 3 | 1 | 4 |
| **Kolkata** | 22.56, 88.36 | 12 | 3 | 1 | 2 | 4 |
| **Badrinath** | 30.74, 79.49 | 3121 | 3 | 2 | 1 | 4 |
| **Delhi** | 28.61, 77.23 | 217 | 3 | 1 | 2 | 4 |
| **Jodhpur** | 26.27, 73.01 | 238 | 3 | 1 | 1 | 4 |
| **Mumbai** | 19.08, 72.88 | 6 | 2 | 1 | 3 | 4 |
| **Nagpur** | 21.15, 79.09 | 311 | 2 | 1 | 2 | 4 |
| **Central Nilgiris** | 11.49, 76.73 | 1843 | 1 | 2 | 3 | 4 |
| **Chennai** | 13.08, 80.27 | 4 | 1 | 3 | 2 | 4 |
| **Ranking total** | | | 23 | 19 | 18 | 38 |

**Table 2:** Weather regime allocations for each weather pattern (from the preferred set of weather patterns) including most common months of occurrence.

| **Weather regime categories** | **Weather patterns associated with each weather regime category** | **Most common months of occurrence (≥ 5%)** |
|---|---|---|
| **Winter Dry Period (WDP)** | 2, 3, 7, 8, 9, 16, 20 | December to March |
| **Western Disturbances (WD)** | 5, 23, 24, 27 | January to May |
| **Pre/Post Summer Monsoon (Pre/Post)** | 12 (mainly pre-monsoon), 13, 14, 15, 22 | May and June (pre-monsoon) and August to October (post-monsoon) |
| **Monsoon Onset (MO)** | 26 | June and July |
| **Active Monsoon (AM)** | 10, 17, 19, 21 | June to September |
| **Break Monsoon (BM)** | 4, 11 | June to August |
| **Retreating Monsoon (RM)** | 1, 6, 18, 25, 28, 29, 30 | September to December |